Extraction of Low-frequent Terms from Domain-specific Texts by Cluster Semantic Analysis

Natalia Archakova, Eugeny Kanevsky
Saint Petersburg Institute for Economics and Mathematics,
Russian Academy of Sciences
Saint-Petersburg, Russia
Assoul@yandex.ru, kanev@emi.nw.ru

Kirill Boyarsky
Saint Petersburg National Research University of
Information Technologies, Mechanics and Optics (ITMO
University)
Saint-Petersburg, Russia
boyarin9@yandex.ru

***Abstract*—We examined a method for extracting the low frequency important single-word terms from domain specific text. Firstly, domain-relevant fragments were extracted from the text with the help of a dependency tree. Then the fragments were clustered and candidate terms were defined using the semantic classifier. The studies suggest that this approach allows extracting even terms with a single occurrence.**

I. INTRODUCTION

Ontology is a set of explicit formal descriptions of terms on a particular domain and of relations between them. Ontology building arouses excessive interest of domain experts; hence, it gets widespread throughout the Web. The aim of these descriptions is to co-operate experts and automated information systems in a given domain. Ontology building is a time-consuming, as it should present a comprehensive notion about concepts and their interconnectivity.

One of the most common approaches to collect data for domain ontology is to use dictionary entries. The terms are selected using linguistic or statistical analysis or hybrid method [1]. A detailed research on applying linguistic and statistic methods can be found in [2]. We adopted a hybrid method as well: nouns are selected with the help of linguistic methods, while candidate-term list is formed as a result of statistical semantic analysis. As in [2], we assessed our results using reference list. However, we focused on single-word terms.

Our technique is rather close to the one used in [3]. The research was based on cluster analysis of English corpus of texts. They applied adaptive Lesk Algorithm “to find the best sense for the two words in each word-to-word pair along with their similarity score”. The results of proposed method were compared with the ones of word frequency, semantic word frequency and position weight methods. Meanwhile even the initial values of Precision and Recall calculated by basic frequency word method were rather high. Though our approaches are quite close, we have different goals. While the article [3] shows how to extract keywords "that can describe the meaning of the document", our research aims to find domain-relevant terms.

A range of techniques is proposed to cluster texts. All of them can be divided into two large groups: hierarchical and non-hierarchical [4]. Each clustering methods has specific limitations. Whatever clustering method would be chosen, a sentence is usually represented in terms of bag-of-words, which means that we disregard the word order and connections between words.

In our study, we explore the ways of building a domain ontology automatically through analyzing terms from economic dictionary. The results of the first stage are presented, including term extraction from automatically detected clusters of terms (Candidate-term list). To give a good understanding to our approach we will discuss three main steps. Firstly, the text was clustered. Then we created a frequency class list for each cluster. The words belonging to the most frequent classes in each cluster are considered candidate terms. At last, the candidate-term list was compared with one determined by experts. The list of terms was bound by single-word nouns (including proper names of organizations and abbreviations).

Relatively small data set allows forming expert term list to estimate the quality of the proposed method.

II. TEXT PRE-PROCESSING

We chose banking domain represented by a part of Russian Great Economical Dictionary [5]. It consists of 1020 entries. Beforehand we used Russian semantic and syntactic parser SemSin [6] for lemmatization, POS-tagging, deep morphological and syntactic disambiguation and partial semantic disambiguation. The output of the analysis is a dependency tree for each sentence. In our research we worked with nouns.

An important part of SemSin is a semantic classifier. As there is no open dictionary of Russian synonyms of high quality (like WordNet for some other languages), we used an update version of the classifier described in [7]. It shows IS-A relation. It has 1633 classes.

The parser dictionary was slightly tuned according to the domain in advance. As a result lexical disambiguation was no more than 1%, which assures robustness of further

ISBN 978-952-68397-6-9 (paperback), ISBN 978-952-68397-7-6 (PDF)

analysis. Preliminary text parsing solves the stop word filter problem. As it was shown in [8], it is preferable to examine nouns in Russian texts comparing to English texts where adjectives and verbs play an important sense role as well. For this reason, we took into account only nouns while other words got into a stopword list.

There are some features caused by the type of this text. Firstly, all entries contain a title and a definition (with a hyperonym) that makes analyzing easier (See Table I). Hereinafter, the terms are converted in upper case.

TABLE I. AN EXAMPLE OF ENTRY

Title		АКЦЕПТНЫЙ ДОМ ACCEPTANCE HOUSE
Definition	**Hypernyme**	Банковское учреждение A banking institution
	Instantiation	специализирующееся на кредитовании внешней торговли. specialized in foreign commerce credit.
	Extra data	АКЦЕПТНЫЙ ДОМ обычно действует на правах акционерной или частной компании. ACCEPTANCE HOUSE usually acts as a stock or private company

On the other hand, terms can be of high and low frequency. The examples of unique terms are РЕДРАФТ (REDRAFT), ХЕДЖЕР (HEDGER). The most frequent word is БАНК (BANK) appearing 826 times in the text. According to [9] such a great range in term frequencies is typical for scientific texts. Standard methods of term extraction (TF-IDF, LDA) often exclude extremum values. A model bag-of-words cannot be fitted either, as it would break inherent structure of a dictionary entry. Thus we explore a part of the dependency tree formed for each dictionary entry instead of applying bag-of-words model. These fragments were constructed with the help of the parser (Fig. 1). Each fragment ("brief entry" in contrast to "full entry") includes a title and its hyponyms expanded with prepositive attributes and their dependent genitive noun. This truncation helps remove most part of general words and find nonadjacent dependencies. The words composing "a brief entry" are bolded below in the example:

ОБМЕННЫЙ КУРС *– **курс**, по которому одна валюта обменивается на другую, **цена денежной единицы страны**, выраженная в иностранной валюте* <…>.

EXCHANGE RATE *– **the rate** at which one currency will be exchanged for another, **the value of another country's currency** compared to that of your own.*

TABLE II. TERM EXTRACTION BY FREQUENCY LISTS CREATED BY DIFFERENT METHODS

Methods	**Word Frequency**		**TF-IDF**		**Proposed method**	
Lemmas	условия	conditions	институт	institute	**заемщик**	**borrower**
	покупка	**sale**	соглашение	agreement	**кредитор**	**lender**
	депозит	**deposit**	условия	conditions	**аккредитив**	**letter of credit**
	прибыль	**profit**	**актив**	**asset**	**цена**	**price**
	средства	funds	**владелец**	**owner**	средства	funds
	требование	requirement	**прибыль**	**profit**	**оплата**	**payment**
	использование	use	уровень	level	**вкладчик**	**depositor**
	производство	production	обращение	treatment	**сделка**	**deal**
	вложение	**investment**	орган	organization	**затрата**	**cost**
	ПОКУПАТЕЛЬ	**CUSTOMER**	**ПОКУПАТЕЛЬ**	**CUSTOMER**	**ПОКУПАТЕЛЬ**	**CUSTOMER**
	расход	consumption	**договор**	**contract**	**компания**	**company**
	соглашение	agreement	**залог**	**pledge**	**облигация**	**bond**
	договор	**contract**	требование	requirement	**залог**	**pledge**
	вкладчик	**depositor**	часть	part	**валюта**	**currency**
	часть	part	**долг**	**debt**	**чек**	**check**
	владелец	**owner**	осуществление	implementation	**фонд**	**fund**
	обращение	treatment	**погашение**	**repayment**	организация	organization
	риск	risk	**чек**	**check**	**документ**	**document**
	время	time	время	time	**рынок**	**market**
	год	year	**заем**	**loan**	расход	expense
Percent of terms	40%		50%		85%	

Besides, some dictionary entries with the same title or different meanings are united into one article. A concept ДИЛИНГ (DEALING) can refer either to a specially equipped location or a provision by financial institutes some services.

To evaluate results of proposed model experts determined 462 single-word terms appearing in the dictionary. The words forming this list will be called terms and marked with upper case. Only 59% of the terms are included in the titles of dictionary entries. 16% of terms appeared in the right part of a"brief entries", including ПЛАТА (PAYMENT), ВКЛАДЧИК (DEPOSITOR), ДЕБЕТ (DEBIT). The other terms can be found only in the "full entries", such as АВАЛЬ (AVAL), БАНКИР (BANKER). Therefore, the analysis of expert term list shows that terms can occur in any part of an entry (title, definition). Hence, in our study we paid more attention to extracting terms from definitions.

To compare the results of our method with standard ones we calculated the word frequency and TF-IDF weights in the whole text and the word frequency of titles. Table II shows fragments of the rank-size distribution of words by different methods. The fragments include for 20 nouns each. The terms are bolded. A term ПОКУПАТЕЛЬ (CUSTOMER) is in the middle of the fragments and it appears 45 times in the given text. Last row shows a percentage of terms in these fragments. The results obtained by frequency (first column) and by TF-IDF weight (second column) are quite poor (40% and 45% correspondingly). This means that we cannot use them for extracting terms.

III. Clustering dictionary entries

Each dictionary entry was represented as a point into a vector space with its values equal to normalized frequencies of token that appear in this dictionary entry. Hence, the sum of normalized frequencies of all tokens in the entry is equal to 1. We use two ways to define tokens. The first option is to take a lemma for a token ("by lemmas" comparing). The second case that we studied is to assign to all lemmas their semantic class ("by classes" comparing). As there is no open dictionary of Russian synonyms of high quality, the classes were assigned according to the classifiers [8]. The class was found after syntactic analysis carried by parser SemSin. That helps to solve a word-sense disambiguation problem. We suggested that lemmas of the same class should have close meanings: for example, words *банкнота* ("banknote") and *валюта* ("currency") belong to the class "Currency notes". Hence, all the measures were computed "by lemmas" and "by classes". Such matrices were formed both for the initial text ("full entries") and for the text of "brief entries".

The clusterization was performed with the help of hierarchical agglomerative Ward algorithm [10]. In many respects, Ward algorithm is considered as the most accurate among other hierarchical methods [13]. Comparing to non-hierarchical methods, it is stable, as it does not depend on initial points. Moreover, it can form clusters of any shape. As stated in [10], in Ward method the bigger cluster the larger inter-cluster distance gets. That allows analyzing texts of low contrast where author segmentation does not correspond to vocabulary changes. Most hierarchical methods prove better results while processing relatively short data sets [4]. Thus since we aim at working with not large corpus, the hierarchical clustering is preferable.

We applied open-source Python package scikit-learn [11] to conduct the clusterization. This implementation is restricted so that the only possible metric to find dissimilarity between points is Euclidean distance. Ward method aims to minimize the change in variance, or the error sum of squares [11, 12].

The optimal clustering "by classes" in terms of inter-cluster and intra-cluster dissimilarity includes 35 clusters containing from 4 to 282 points. The best result conducted "by lemmas" consists of 35 clusters. The clusters include from 4 to 436 points.

TABLE III. Average intra-cluster and inter-cluster distances

Text	Intra-cluster distance	Inter-cluster distance
"by classes"	0.41	1.04
"by lemmas"	0.45	1

Both results ("by lemmas" and "by clusters") were compared through calculating inter-cluster and inter-cluster distances (Table III). The higher the difference between them, the more accurate clusterization was conducted. Here it ranges in 0..1.41.

Generally, word selection "by classes" is not worse than "by lemmas" selection. It shows higher results in finding middle-size clusters suitable for further analysis. For example, a term ЦЕССИОНАРИЙ (CESSIONARY) has three meanings: a person, who becomes a creditor (1); a legal successor (2); as insurance company (3). In all variants meaning (1) corresponds to a cluster "*People*". Meaning (2) belongs to the cluster "*People*" and meaning (3) gets into the cluster "*Financial institutes*" while selecting "by classes". In the other case ("by lemmas") meanings (2) and (3) are in the largest cluster.

IV. Evaluation and Discussion

The final stage includes two steps. The clusterization was produced on the base of "brief entry" to obtain clear results. The final candidate term lists were formed using the "full entries" to extract as much relevant words as it is possible.

On the first step, we formed a list of most frequent classes from each cluster. It has 36 classes. Among them, there are "Finance", "Money", "Payment", "Institutes". Comparing to the 36 most frequent classes chosen before clustering, a proposed technique helps find 7 new valuable classes: for example, "Documents" and "Trade and Service" with terms MARKET and AUCTION.

On the second step, we formed the candidate term list that includes words of these classes. For example, a word *bank* would be a candidate-term only if its class is frequent in the same cluster where this word appears.

A class term list includes 311 candidate terms of which 249 (80%) are real terms. Comparing to the example in sec. III (table II) a given subset of the candidate-term list (the third

column) has 13 terms that do not appear in other columns, for example *BORROWER, LENDER, CURRENCY, FUND, MARKET*. At the same time, the first and the second columns presented standard methods have only 3 unique terms each (*sale, deposit, investment, asset, debt, repayment*).

Deep lexical analysis of candidate terms shows that there were found 95 of 147 terms that appear once: PREPAYMENT, LOUIS (a coin), FIDUCIARY, HOLDING COMPANY. Among them 25 terms occur only in "full entries" and 15 terms appear in the definitions of the "brief entries".

As evaluation metrics, we chose the Precision and Recall, which are considered to be the standard metrics for retrieval effectiveness in informational retrieval. Precision is a part of expert terms found automatically, while Recall is a part of candidate terms in terms extracted by expert.

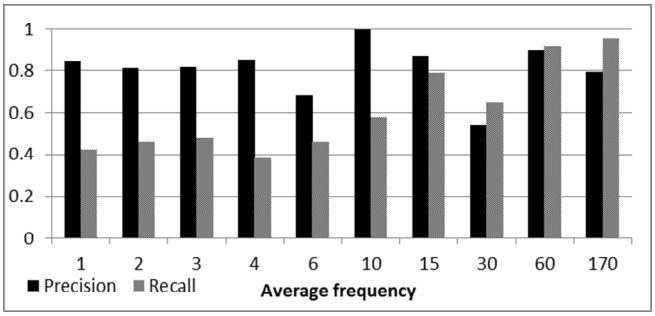

Fig. 1. Precision and Recall of automatically extracted terms

Fig. 1 shows the evaluation of Precision and Recall for 10 equal intervals so that terms with the same frequency get into the same interval. The average frequency for each interval is rounded. The precision does not depend on the term frequency. At the same time, the more frequent a term, the higher probability that the method would find all terms with this frequency. This method can extract terms occurred once and twice with the probability of 40%. It is worth mentioning that standard methods cannot extract terms with such a low frequency.

TABLE IV. PRECISION AND RECALL FOR TERM EXTRACTION

	Precision	**Recall**
Word Frequency	0.37	0.25
TF-IDF	0.27	0.19
Title word frequency	0.63	0.42
Proposed approach "by lemmas"	0.64	0.48
Proposed approach "by classes"	0.8	0.54

Then we decided to compare average precision and recall calculated by four different methods: by word frequent list of the full text, by TF-IDF weights, by word frequent list of titles of the dictionary and by the proposed approach.

We chose 311 first most frequent words from each list as a list of candidate terms has the same size. Note that the 311th word from the word frequent list occurs only 10 times. Each list was compared to the expert term list.

V. CONCLUSION

In this research, we presented a multistage method for extracting the low frequency important single-word terms from domain specific text. To do so, we used a domain dictionary, the classifier similar to the WordNet and the semantic and syntactic parser.

The experimental results suggest that standard techniques of word frequency and TF-IDF weights cannot show a real picture to term distribution through text. Besides, we found that terms can occur in any part of the dictionary entry, thus we could not confine our research to a simple analysis of entry's titles. The quality of automatic term extraction was estimated with the help of expert term list.

The proposed method almost doubles the probability of term extraction. It does not depend on an initial word frequency in the text though the clustering results should be of a high quality. Eventually, the candidate term list includes terms of any frequency. A final F-score of candidate terms is 65% with average precision of 80% and recall of 54%.

REFERENCES

[1] J. Lacasta, J. Nogueras-Iso, and J. Zarazaga-Soria, *Terminological Ontologies: Design, Management and Practical Applications.* Semantic Web and Beyond: Computing for Human Experience. Springer-Verlag, 2010.

[2] M.T.Pazienza, M. Pennacchiotti, and F.M. Zanzotto, "Terminology extraction: an analysis of linguistic and statistical approaches". *Knowledge Mining,* Springer Verlag, 2005, pp 255-281.

[3] M.H. Haggag, A. Abutabl, and A. Basil, "Keyword extraction using Clustering and Semantic Analysis", *International Journal of Science and Research*, vol.3, Nov.2014, pp. 1128 -1132.

[4] M. Steinbach, G. Karypis, V. Kumar. A Comparison of Document Clustering Techniques. Web: http://www-users.cs.umn.edu/~karypis/

[5] A.B. Borisov, *Great Dictionary of Economics.* Moscow: Knizhny mir (Word of Books), 2003.

[6] K.K. Boyarsky, and E.A. Kanevsky, "Semantics and syntactics parser SemSin", *Scientific and technical journal of information technologies, mechanics and optics*, vol.15, May 2015, pp. 869-879.

[7] V.A. Tuzov, Computer semantics of the Russian language. Saint-Petersburg: publishing house of St. Petersburg University. University Press, 2004.

[8] N. Avdeeva, G. Artemova , K. Boyarsky, N. Gusarova, N. Dobrenko, E. Kanevsky, "Subtopic Segmentation of Scientific Texts: Parametr Optimisation", *Communications in Computer and Information Science*, vol.518, 2015, pp. 3–15.

[9] G. Artemova , K. Boyarsky, D. Gouzévitch, N. Gusarova, N. Dobrenko, E. Kanevsky, and D. Petrova, "Text Categorization for Generation of Historical Shipbuilding Ontology", *Communications in Computer and Information Science*, vol.468, 2015, pp. 1–14.

[10] R. Srinivasan, A.Pepe, and V.F. Rodrigez, "A Clustering-Based Semi-Automated Technique to Build Cultural Ontologies", Journal of the American Society for Information Science and Technology, vol.10, Feb.2009, pp. 1-13.

[11] SciPy Web: http://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html#scipy.cluster.hierarhie.linkage.

[12] K.V. Vorontsov, Lectures on clustering algorithms and multidimensional scaling, 2007. Web: http://www.ccas.ru/voron/download/Clustering.pdf